Revised Data Metrics for 2020 Disclosure Avoidance

Update for the June 8, 2021, Production Settings Version Metrics Release

This release reflects the parameters chosen by the Data Stewardship Executive Policy Committee (DSEP) on June 8, 2021 for the production run of the 2020 Census Disclosure Avoidance System. This is the fifth release of detailed summary metrics that allow data users external to the Census Bureau to assess the 2020 Disclosure Avoidance System (DAS). This release is not accompanied by a new set of privacy-protected microdata files (PPMFs) at this time; the Census Bureau plans to release a set of Production Settings PPMFs in September. The five released vintages are as follows:

(1) 2010 Summary Metrics and Demonstration PPMF Version 2020-05-27
(2) 2010 Summary Metrics and Demonstration PPMF Version 2020-09-17
(3) 2010 Summary Metrics and Demonstration PPMF Version 2020-11-16
(4) 2010 Summary Metrics and Demonstration PPMF Version 2021-04-28
(5) 2010 Summary Metrics Production Settings (no PPMF)

The detailed summary metrics contain measurements of accuracy, bias, and outliers for the person data and the unit data, and they also measure the level of consistency between the two files.

While metrics have been developed for the full complement of 2020 Census variables, the variables included in metrics releases may be limited based on operational priorities. The June 2021 Production Settings run only included the variables necessary to create the redistricting (P.L. 94-171) data; tables are left blank when the data necessary to produce the metric are not available in the Production Settings data.

We welcome feedback and questions on this document. Please submit feedback on these revised metrics to: 2020DAS@census.gov.

Background

The Census Bureau is developing a new method of disclosure avoidance for the 2020 Census (referred to as the top-down algorithm) to protect the privacy of respondents. In October 2019, the Bureau released a set of protected tabulations based on 2010 Census responses, known as the 2010 Demonstration Data Products, to show data users how this new disclosure avoidance system might impact the accuracy of data products.

Data users gave feedback on the demonstration products to the Census Bureau both by email and at a workshop hosted by the National Academy of Sciences Committee on National Statistics in December 2019. Much of the feedback focused on concerns regarding the accuracy of the post-disclosure protected tabulations (i.e., how close the new tabulations were to the original tabulations) and bias (i.e., whether the new tabulations systematically differed from the original tabulations due to population size or other characteristics).

There are two core components to the 2020 DAS: noise injection and post-processing. In order to protect privacy, the DAS injects a small amount of noise into every statistic that it produces from the confidential data. The amount of noise for each statistic is randomly selected from a distribution centered around zero. Post-processing ensures that the records add up across geographies and do not include negative values. Distinguishing between the portion of total error that is attributable to post-processing and the error attributable to the injection of noise into the data is an area of interest to data users. However, this is outside the scope of the design of these metrics which are intended to provide measures of the total error. Data users also highlighted specific geographies for inclusion in the metrics, including: counties, political entities such as incorporated places, and American Indian/Alaska Native/Native Hawaiian (AIANNH) Areas.

This document describes a series of metrics developed to assess both the 2010 Demonstration Data Products and future development runs of the DAS, as improvements are made leading up to the release of 2020 Census data products. As testing and development of the disclosure avoidance system continues, these metrics will be used to concisely and quantitatively communicate data quality improvements to data users and the broader stakeholder community.

The intent is not to replicate a full analysis of each development run, but to provide a set of metrics that will inform stakeholders of the fitness of use across variables and geographies. These metrics show the accuracy of both a broad set of demographic measures and specific types of use cases. The included metrics, and the formulation of metrics for specific use cases, will evolve and new metrics may be added based on external feedback.

This document contains examples for the resident population of the United States. The resident population of Puerto Rico will be analyzed in a similar manner; however, statistics for the United States will not be pooled with statistics for Puerto Rico.

Metrics

Based on the feedback from the 2010 Demonstration Data Products, data users are concerned about the impact of the new disclosure avoidance methodology on accuracy, bias, outliers, and impossible or improbable results.

In order to be able to assess the impact of the new disclosure avoidance methodology, the metrics will provide information about the change from an initial set of tabulations, prior to the application of the new disclosure avoidance methodology, and the same tabulation after the application of the new disclosure avoidance methodology. These metrics will be produced using tabulations developed from the 2010 Demonstration Data Product Microdata Detail File (MDF) and subsequent runs of the disclosure avoidance system using the new disclosure avoidance methodology – referred to as “MDF.” The MDF is generated by applying the new disclosure avoidance methodology to the 2010 Census Edited File (CEF), an internal 2010 Census file that has not been protected using the 2010 disclosure avoidance methodology known as “swapping”. The publicly available 2010 Census tabulations that the MDF tabulations are compared to are from the 2010 Census Hundred-percent Detail File (HDF) that has been protected using the 2010 Census disclosure avoidance methodology (swapping). All publicly released analysis will be done based on comparisons between tabulations from the 2010 MDF and the 2010 Census HDF. In the formulas below, MDF means “tabulated from the Microdata Detail File” and CEF means “tabulated from the Census Edited File.” Most of the comparisons that the Census Bureau will present through the metrics, and all of the comparisons that were done by external users of the 2010 Demonstration Data Products, substitute HDF for CEF in these formulas, meaning “tabulated from the

Hundred-percent Detail File (swapped data)." The conceptually correct error measure is relative to the CEF, but in order to document the issues raised by external reviewers, the first collection of values for these metrics was based on the HDF so that external users could verify that the Census Bureau had implemented the metric correctly. For consistency, future values for metrics will also reflect differences between the MDF and HDF.

Accuracy

Accuracy is measured by comparing the MDF to the CEF. Accuracy can be "absolute" or "relative" – that is, accuracy can be measured as either a count (the total population differed by 20 people) or as a percent of the original (the total population differed by 5%).

The following metrics for accuracy will be used:

1. **Mean Absolute Error (MAE)**: This is a measure of the "average" absolute value of the count difference for a particular statistic. For example, for total population at the county level, calculate Abs(MDF – CEF) for each of the 3,143 counties, then take the mean.[1]
2. **Mean Numeric Error (ME)**: This is a measure of the magnitude and direction of the average difference for a particular statistic. For example, for total population at the county level, calculate (MDF – CEF) for each of the 3,143 counties, then take the mean.
3. **Root Mean Squared Error (RMSE):** This is a measure of the square root of the average squared error for a particular statistics. It is the traditional measure of error for Census Bureau sample survey statistics. For example, for total population at the county level, calculate (MDF – CEF)^2 for each of the 3,143 counties, take the mean, then take the square root.
4. **Mean Absolute Percent Error (MAPE)**: This is a measure of the "average" relative difference for a particular statistic. For example, for total population at the county level, calculate [Abs(MDF – CEF)/CEF] for each of the 3,143 counties, then take the mean.
5. **Coefficient of Variation (CV):** This is the relative error counterpart to RMSE. It is another traditional measure of error in Census Bureau sample survey statistics. For the same collection of statistics as was used for RMSE, calculate Avg(CEF), then calculate [RMSE/Avg(CEF)].
6. **Total Absolute Error of Shares (TAES):** This measure finds the proportion of each MDF value to the total MDF value for the summary geography and subtracts the proportion of the CEF value to the total CEF value for the summary geography. The absolute value of these proportional differences across evaluation geographies is then summed to the summary geography level. The goal is to provide a measure of the distributional error in the MDF shares.
7. **Percent Difference Thresholds = Count of absolute percent differences above a certain threshold:** Unlike the other measures, Percent Difference Thresholds are a numeric value that rely upon a set threshold (e.g., 5 and 10 percent). In short, the absolute percent difference is computed by dividing the absolute difference between the MDF and CEF value for a given geography by the CEF value for that geography and multiplying by 100. The end measure simply represents a count of how many evaluation geographies exceed a particular threshold in their absolute percent difference of the estimate. It provides a measure of the distribution of differences.

[1] The reference to "counties" includes counties and county equivalents in the 2010 Census – the list of counties in the 2010 Census is located here: https://www.census.gov/geographies/reference-files/time-series/geo/tallies.html

Accuracy will be calculated using the above metrics both overall (e.g., for all 3,143 counties) and also for particular population and cell size categories (e.g., for counties with populations below 10,000 people or cells with counts equal to or greater than 100). For tracts, the MAE and RMSE are used as the primary error measures for determining accuracy. This is because tracts are roughly equal in size, so the magnitude and direction of the change is less important.

Bias

Bias is a concept related to accuracy, but direction of change and whether that varies with population size or other characteristics is what matters most. Prior research into the top-down algorithm (TDA) post-processing has demonstrated that with early versions of the TDA geographic areas with small populations (or statistics with small cell sizes) tend to have a positive bias - where the MDF tabulation is systematically greater than the CEF tabulation, while those areas with larger populations (or larger cell sizes) tend to have a negative bias.

The following metrics for bias will be used:

1. **Mean Numeric Error (ME)**: This is a measure of the magnitude and direction of the average difference for a particular statistic. For example, for total population at the county level, calculate (MDF – CEF) for each of the 3,143 counties, then take the mean.
2. **Mean Percent Error (MALPE)**: This is a measure of the magnitude and direction of the average relative difference for a particular statistic. For example, for total population at the county level, calculate [(MDF – CEF)/CEF] for each of the 3,143 counties, then take the mean.

Bias will generally be calculated by population size or cell size categories (e.g., categories for counties below 1,000 people, counties between 1,000 to 4,999 people, counties between 5,000 to 9,999 people, counties between 10,000 and 49,999 people, counties between 50,000 and 99,999 people, and counties equal to or greater than 100,000 people). [2] Bias will also be calculated by urban/rural classification. Urban areas will be classified based on the Census Bureau's 2010 classification that require them to be comprised of a densely settled core of census tracts and/or census blocks that meet minimum population density requirements, along with adjacent territory containing non-residential urban land uses as well as territory with low population density included to link outlying densely settled territory with the densely settled core.[3] "Rural areas" encompass all population, housing, and territory not included within an urban area. Using the metrics proposed above, the amount of bias introduced to urban and rural areas will be calculated.

For certain statistics and geographic areas, the distribution of proportional differences across subordinate geographies matters greatly. The metric **Total Absolute Error of Shares (TAES)** is proposed to measure how close the disclosure-protected spatial distribution is to the 2010 Census internal data distribution. It is calculated as follows: $\sum_i \left|\frac{MDF_i}{\sum_i MDF_i} - \frac{CEF_i}{\sum_i CEF_i}\right|$, where MDF_i is an individual subordinate geography's privatized tabulated value and CEF_i is an individual subordinate geography's 2010 Census value. To illustrate, imagine a county with two tracts: one that contains 90 percent of the county's

[2] Size categories were evaluated to determine best fit and may be adjusted.

[3] To qualify as an urban area, the territory must encompass at least 2,500 people with at least 1,500 residing outside of institutional group quarters. The Census Bureau identifies two types of urban areas: Urbanized Areas (UAs) of 50,000 or more people and Urban Clusters (UCs) of at least 2,500 and less than 50,000 people.

population and one that contains the other 10 percent. If the privatized data now have equal populations in each tract for a hypothetical county, the TAES will be calculated as [Abs(0.5 - 0.9) + Abs(0.5 - 0.1)] = 0.8.

Additional information related to the equations used to compute the metrics is located in the Appendix.

Outliers and Impossible or Improbable Results

Additionally, certain statistics will be internally examined for "outliers": What is the largest increase in tabulated value? What is the largest decrease? Is there an inconsistency across the person and unit tables that is impossible or highly improbable? These will inform internal evaluations about the plausibility of tabulated results. Counts of outliers will be made available externally to allow for an assessment of the number of entities with exceptionally large differences between the MDF and the CEF for several of the data metrics tables.

Geographic Levels

Based on feedback received from the 2010 Demonstration Data Products, data users are particularly concerned about data fitness for states, counties (including county equivalents), political entities such as incorporated places or minor civil divisions (MCDs), Federal American Indian Reservations/Off-Reservation Trust Lands, Oklahoma Tribal Statistical Areas (OTSAs), Alaska Native Village Statistical Areas (ANVSAs), and, for limited use cases, tracts, block groups, and blocks. Additional sets of metrics will be provided for Puerto Rico municipios and tracts, as well as additional levels of geography such as Elementary, Secondary, and Unified School Districts.

The metrics presented here are generally national level aggregations of lower levels of geography.

Use Cases and Proposed Metrics

The metrics include an extensive set of general measures that provide an accuracy profile for each DAS development run. This accuracy profile will provide information on the fitness of use for many critical uses.

Additional metrics were developed for specific categories of use cases. Use cases were identified through a Federal Register Notice, the Committee on National Statistics (CNSTAT) Demonstration Products Workshop, and other outreach. The categories were created based on the type of accuracy that was the most important for the use cases within that category. While several accuracy measures are provided, each category has a primary measure for assessing fitness of use. This allowed for metrics to be developed that were designed specifically for the following categories of use cases:

Zero-Sum Total: Uses that rely on the accuracy of the distribution in addition to the overall accuracy because a fixed amount of something is being distributed across categories. For these uses, the accuracy needs may be greater for the distribution than for the actual estimates. For these types of use cases, the TAES would serve as the primary measure for fitness of use.

Zero-Sum Category: Same as zero-sum total except use cases rely on estimates for some subset of the total. For these types of use cases, the TAES would serve as the primary measure for fitness of use.

Variable-Sum Total: Similar to zero-sum use cases except that the total of what is being distributed can vary. For variable-sum total, the accuracy of the estimate is more important than the accuracy of the distribution. For these types of use cases, the MAPE would serve as the primary measure for fitness of use.

Variable-Sum Category: Same as variable-sum total but for a subset of the population. For these types of use cases, the MAPE would serve as the primary measure for fitness of use.

Single Year of Age Accuracy: These use cases require accuracy for single years of age rather than age groups. For these types of use cases, the MAPE would serve as the primary measure for fitness of use.

Rates Accuracy: These uses cases rely on a measure of the size of a subgroup(s) within the total population. For these types of use cases, because they are based on a rate, the MAE and RMSE as a percentage point difference serves as the primary measure for fitness of use.

Percent Threshold: Use case depends on the subset of the population crossing a percent threshold. For these types of use cases, counts of geographic entities crossing the threshold would serve as the primary measure for fitness of use.

Numeric Threshold: Use case depends on the subset of the population crossing a numeric threshold. For these types of use cases, counts of geographic entities crossing the threshold would serve as the primary measure for fitness of use.

Basic Demographic Accuracy Profile

The content in the Basic Demographic Accuracy Profile is consistent with the demographic characteristics expected in the 2020 Census national redistricting data, with the exception of tables 15, 16, 17 and 19.

Total Population

Total population at the state level is invariant so a measure of accuracy is not needed. Measures of change in the Total Population and Total Population 18 Years and Over will be provided for multiple levels of geography. [Tables 1a-k and 2a-i]

MAE, RMSE, MAPE, CV, and MALPE will be provided for Total Population for the following geographies:

- Counties by size categories
- Incorporated places by size categories
- Urban/rural based on the block-level urban/rural designation
- Puerto Rico Municipios
- Elementary, Secondary, and Unified School Districts by size categories
- MCDs by size categories
- Federal American Indian Reservations/Off-Reservation Trust Lands by size categories
- OTSAs
- ANVSAs by size categories

MAE and RMSE will be provided for Total Population for the following geographies:

- Tracts
- Puerto Rico Tracts

MAE, RMSE, MAPE, CV, and MALPE will be provided for Total Population 18 Years and Over for the following geographies:

- Counties by size categories
- Incorporated places by size categories
- Urban/rural based on the block-level urban/rural designation
- Puerto Rico Municipios
- MCDs by size categories
- Federal American Indian Reservations/Off-Reservation Trust Lands by size categories
- OTSAs
- ANVSAs by size categories

MAE and RMSE will be provided for Total Population 18 Years and Over for the following geographies:

- Tracts
- Puerto Rico Tracts

Size categories for counties, school districts, and MCDs are:

- Less than 1,000 people
- 1,000 to 4,999 people
- 5,000 to 9,999 people
- 10,000 to 49,999 people
- 50,000 to 99,999 people
- 100,000 people or more

Size categories for incorporated places are:

- Less than 500 people
- 500 to 999 people
- 1,000 to 4,999 people
- 5,000 to 9,999 people
- 10,000 to 49,999 people
- 50,000 to 99,999 people
- 100,000 people or more

Size categories for Federal American Indian Reservations/Off-Reservation Trust Lands and ANVSAs are:

- Less than 100 people
- 100 to 999 people
- 1,000 to 9,999 people
- 10,000 people or more

Because of the standard size of tracts, the tract-level measures will not be provided by size categories.

For all geographies, secondary measures of outliers will be provided. This measure will include counts of geographies where the absolute percent difference is “2 to 5 percent” or “exceeds 5 percent.” An additional outlier measure will provide the counts of geographies where the absolute numeric difference exceeds 200.

In the previous version of these metrics, error measures were provided for the non-Hispanic white population. Based on the expansion of the race and Hispanic origin metrics, the table showing the total population by percent of population that is non-Hispanic white was removed.

Total Housing Units

Counts of housing units are invariant at the block level; therefore a measure of accuracy is not needed.

Occupancy and Households

Measures of change in the occupancy rate and persons per household will be provided for multiple levels of geography. [Tables 3a-j and 4a-e]

Because occupancy is expressed as a rate, the MAE, RMSE, and MALPE is modified to reflect the percentage point difference. **Modified MAE - mean absolute percentage point error, RMSE, and the modified ME- mean percentage point error** will be provided for the occupancy rate for the following geographies:

- Counties
- Incorporated places
- Puerto Rico Municipios
- Elementary, Secondary, and Unified School Districts
- MCDs
- Federal American Indian Reservations/Off-Reservation Trust Lands
- OTSAs
- ANVSAs

Modified MAE - mean absolute percentage point error and RMSE will be provided for the occupancy rate for the following geographies:

- Tracts
- Puerto Rico Tracts

For the occupancy rate, a secondary measure of outliers will be provided for all geographies: counts of where the occupancy is 100 percent in the MDF but not the CEF, and where the occupancy is 0 percent in the MDF but not in the CEF. Counts of where the error of occupancy rate is “2 percentage point to 5 percentage points” or “exceeds 5 percentage points” will also be provided.

Persons-per-household is derived by dividing the household population by the number of households. **MAE, RMSE and ME** will be provided for persons per household for the following geographies:

- Counties by size categories
- Incorporated place size categories
- Urban/rural based on the block-level urban/rural designation.
- Puerto Rico Municipios

MAE and RMSE will be provided for persons per household for the following geographies:

- Tracts
- Puerto Rico Tracts

For persons per households, a secondary measure of outliers for all geographies will be provided. This measure will include counts of geographies where the absolute percent difference is “2 to 5 percent” or “exceeds 5 percent.”

In the previous version of these metrics, error measures were provided showing the count of tracts where the population total is less than the population derived from the household size variable. This table has been moved to the new section for Impossible and Improbable Results Use Cases.

Race and Hispanic Origin

Error measures will be provided for several Hispanic origin and race groupings.

MAE, RMSE, MAPE, CV, and MALPE will be provided for “Hispanic or Latino Origin” [Table 5a-c], “Hispanic or Latino Origin by Race Alone” [Table 8a-c], “Hispanic or Latino Origin by Race alone or in combination with one or more other races” [Table 9a-c], “Number of Races” [Table 10a-c], and “Hispanic or Latino Origin by Number of Races” [Table 11a-c] for the following geographies:

- All states
- Counties by size categories
- Incorporated places by size categories

MAE and RMSE will be provided for “Hispanic or Latino Origin” [Table 5d], “Hispanic or Latino Origin by Race Alone” [Table 8d], “Hispanic or Latino Origin by Race alone or in combination with one or more other races” [Table 9d], “Number of Races” [Table 10d], and “Hispanic or Latino Origin by Number of Races” [Table 11d] for the following geographies:

- Tracts by size categories

MAE, RMSE, MAPE, CV, and MALPE will be provided for “Race Alone” [Table 6a-i] and for “Race alone or in combination with one or more other races” [7a-i] for the following geographies:

- All states
- Counties by size categories
- Incorporated places by size categories
- Puerto Rico Municipios by size categories
- MCDs by size categories
- Federal American Indian Reservations/Off-Reservation Trust Lands by size categories
- OTSAs by size categories
- ANVSAs by size categories

MAE and RMSE will be provided for “Race Alone” and for “Race alone or in combination with one or more other races” for the following geographies:

- Tracts by size categories
- Puerto Rico Tracts by size categories

Size categories for all geographies, except states, for these tables are:

- Population between 0 and 9 for the race/Hispanic origin category
- Population between 10 and 99 for the race/Hispanic origin category
- Population of 100 or more for the race/Hispanic origin category

To supplement analyses conducted by other areas for the redistricting data product, **MAE and RMSE** will be provided for following Hispanic origin and race groupings by voting-age population (18 years and older) at the tract and block group levels:

- Hispanic or Latino Origin by Race Alone for the Population 18 Years and Over [Table 12a-b]
- Hispanic or Latino Origin by Race alone or in combination with one or more other races for the Population 18 Years and Over [Table 13a-b]
- Hispanic or Latino Origin by Number of Races for the Population 18 Years and Over [Table 14a-b]

An additional outlier metric will be provided for all geographies to show where the absolute percent difference exceeds 10%.

Age and Sex

Measures of accuracy for age and sex will be provided for multiple age groupings.

MAE, RMSE, MAPE, CV, and MALPE will be provided for Sex by Ages 0-17, 18-64, and 65 and over [Table 15a-b] for the following geographies:

- Counties by limited size categories
- Incorporated places by limited size categories

Size categories for counties and incorporated places are:

- All counties/incorporated places
- Less than 1,000 people

MAE and RMSE will be provided for Sex by Ages 0-17, 18-64, and 65 and over [Table 15c] for the following geographies:

- Tracts

MAE, RMSE, MAPE, CV, and MALPE will be provided for Sex by Age in 5-year age bins from 0-115 [Table 16a-h] for the following geographies:

- Counties
- Incorporated places
- Puerto Rico Municipios
- Federal American Indian Reservations/Off-Reservation Trust Lands
- OTSAs
- ANVSAs

MAE and RMSE will be provided for Sex by Age in 5-year age bins from 0-115 for the following geographies:

- Tracts

- Puerto Rico Tracts

An additional outlier metric will be provided for all geographies to show where the absolute percent difference exceeds 10%.

A new table has been added to show the average absolute change in the sex ratio and median age for county size categories [Table 17]

Size categories for counties in Table 17 are:

- Less than 1,000 people
- 1,000 to 4,999 people
- 5,000 to 9,999 people
- 10,000 to 49,999 people
- 50,000 to 99,999 people
- 100,000 people or more.

Group Quarters Population by Major GQ Type and Institutionalized versus Noninstitutionalized
Measures of accuracy for the population in group quarters will be provided by major group quarters type. The universe for these tables is restricted to those with at least one group quarters.

Major GQ Types are classified as:

- **Institutional Group Quarters:** 1) Correctional Facilities for Adults, 2) Juvenile Facilities, 3) Nursing Facilities/Skilled-Nursing Facilities, 4) Other Institutional Facilities
- **Noninstitutional Group Quarters**: 5) College/University Student Housing, 6) Military Quarters, 7) Other Noninstitutional Facilities

MAE, RMSE, MAPE, CV, and MALPE will be provided for Group Quarters Population by Major GQ Type and Noninstitutionalized [Tables18a-c] for the following geographies:

- States
- Counties by size categories
- Incorporated places by size categories

MAE and RMSE will be provided for Group Quarters Population by Major GQ Type and Noninstitutionalized [Tables18d] for the following geographies:

- Tracts

Size categories for counties are:

- Less than 1,000 people
- 1,000 to 4,999 people
- 5,000 to 9,999 people
- 10,000 to 49,999 people
- 50,000 to 99,999 people
- 100,000 people or more

Size categories for incorporated places are:

- Less than 500 people
- 500 to 999 people
- 1,000 to 4,999 people
- 5,000 to 9,999 people
- 10,000 to 49,999 people
- 50,000 to 99,999 people
- 100,000 people or more

An additional outlier metric will be provided for all geographies to show where the absolute percent difference exceeds 10%.

A new table has been added to show the average absolute change in the sex ratio and median age for the group quarters population by county size categories [Table 19].

Size categories for counties in Table 19 are:

- Less than 1,000 people
- 1,000 to 4,999 people
- 5,000 to 9,999 people
- 10,000 to 49,999 people
- 50,000 to 99,999 people
- 100,000 people or more.

<u>Categories of Use Cases with Specific Examples</u>

Emergency Service Planning for a Specific Population within a Small Geographic Area

Variable-sum category (local)

A specific example of this type of use case is a scenario where the number of people aged 75 and over is required to determine the number of buses or other resources needed to evacuate the elderly population from an area. This type of use case is representative of a local, non-zero-sum category use case since the number of buses is not limited and will be based on the size of the population in need. This makes the size of the target population the population measure that requires accuracy. There is also a geographic need, since the buses would need to be staged in close vicinity to the population in need. This type of use case tends to be for smaller geographic areas and most often requires counts of the elderly or of children.

MAE and RMSE will be the primary measure of error for the population aged 75 years and over [Use Case Table 1] for the following geographies:

- Tracts

Counts of the tracts where the absolute percent difference is 2% to 5% and where the absolute percent difference exceeds 5% for the target population group will be provided as a secondary measure of fitness for use. [Use Case Table 1]

In the previous version of these metrics, error measures were provided for young children (under 5 years of age) as a use case table. Based on the inclusion of these error measures for the population under 5 years of age in the Basic Accuracy Profile, this table was removed.

Distribution of Federal Funds

Zero-Sum Total

The distribution of federal funds use case is generally understood to be a state, county, incorporated place, and MCD level distribution of a fixed amount. Because state-level counts are invariant, a state level measure isn't needed. With this type of use case, a fixed amount is distributed based on each area's share of the population, making the accuracy of the shares, or the distribution, the primary measure that requires accuracy.

The primary measure to assess fitness of use for this use case will be the **TAES** at the county level within each state as a share of that state, at the incorporated place level as a share of that state and at the MCD level as a share of that state. [Use Case Tables 2, 3, and 4]

A table showing TAES for the total population at the county level as a share of the nation was removed from the revised version of the metrics. As a result of the total population at the state level being invariant, this measure did not yield useful information for external data users.

Projections of the Population Entering School or Eligibility for a Program

Single Year of Age Accuracy

This use case requires accuracy for counts of people of a single year of age or age ranges. For this type of use case, single year of age accuracy may be needed for a single year of age or for an age range, for example, those entering school, or those who will be graduating school, or those who will be eligible for different programs for a set number of years in the future. Other examples include those expected to complete immunization schedules; expected draft registration; eligibility for retirement, Medicare, or Social Security; or, more broadly, projections of the population by single year of age.

The measures of accuracy for assessing fitness of use for this use case are the same as for the total population, but applied to a specific age or age range. The accuracy need is in the counts of the population in the specified age or age range.

MAE, RMSE, MAPE, CV, and MALPE will be the primary measure of error for single years of age 0 through 17 for the following geographies:

- Counties by size categories
- Elementary, Secondary, and Unified School Districts by size categories

Size categories for counties and school districts are:

- less than 1,000 people under 18 years old
- 1,000 to 9,999 people under 18 years old
- 10,000 people or more under 18 years old

A secondary measure will be provided for outliers, which will be the count of counties and Elementary, Secondary, and Unified School Districts where the absolute percent difference exceeds 5 percent. [Use Case Table 5a-d]

Previous versions of these metrics only provided information for ages 4 and 17 for the county, incorporated place, and tract level. A table showing the TAES for the share of counties and places within the nation for these two ages was also provided. Based on feedback from external stakeholders, the above measures and geographies for single year of age were determined to best meet stakeholder needs and the other tables were removed.

Total Population for American Indian and Alaska Native Race Groups

Zero- and Variable-Sum Category

Federal funding use allocation formulas such as the Tribal Transportation Programs and Indian Housing Block Grant funding rely on Census data. These uses require accuracy of the counts of the American Indian and Alaska Native population.

The measure of accuracy used for this use case will be the **TAES.** The **TAES** measure will be applied to the AIAN population distribution across counties and incorporated places within the nation. [Use Case Table 6]

In the previous version of these metrics, error measures were provided specifically for the AIAN population alone or in combination with one or more other races. Because the measures for Race alone or in combination were added in the Basic Accuracy Profile, this table was removed.

Outreach for Rare/Small Populations – Race Use Cases

Variable-Sum Total

This use case depends on the accuracy of the data for measuring rare or small populations – these metrics will focus on how accurately the presence of AIAN and NHPI alone populations can be determined. Fitness of use depends on being able to correctly identify small populations with a minimal number of false positives or false negatives, or the ability to show when a population exists in an area, and when it does not exist.

An outlier metric will be available that shows the counts of where AIAN alone population in the MDF is less than in the CEF, and the median percentage of reduction for these areas will be provided for the following geographies [Use Case Table 7a-e]:

- Counties by size categories
- Incorporated places by size categories
- Federal American Indian Reservations/Off-Reservation Trust Lands by size categories
- OTSAs by size categories
- ANVSAs by size categories

Size categories for all geographies are:

- between 0 and 9 people who are AIAN alone
- between 10 and 99 people who are AIAN alone
- 100 or more people who are AIAN alone

An outlier metric will also be available that shows the counts of where NHPI alone population in the MDF is less than in the CEF, and the median percentage of reduction for these areas, are provided for the following geographies [Use Case Table 8a-b]:

- Counties by size categories
- Incorporated places by size categories

Size categories for all geographies are:

- between 0 and 9 people who are NHPI alone
- between 10 and 99 people who are NHPI alone
- 100 or more people who are NHPI alone

A secondary measure of fitness for use will be provided to identify clusters of AIAN and NHPI population in tracts, with the minimum population to indicate a cluster being the presence of at least 100 people in a tract that are either AIAN (alone or in combination) or NHPI (alone or in combination). A count of false negatives and false positives will be provided for tracts. A false positive will be defined as when the CEF population is equal to or greater than 100 and the MDF population is less than 20. A false negative will be defined as when the CEF population is less than 20 and the MDF population is equal to or greater than 100. [Use Case Table 9-10]

In the previous version of these metrics, error measures were provided specifically for the AIAN population alone or in combination with one or more other races. Because the measures for Race alone or in combination were added in the Basic Accuracy Profile, this table was removed.

Target Vacancy/Occupancy Rates

Percent/Rate Thresholds

In this use case, a threshold has been established as a target or as a threshold for inclusion. A specific example is the use of vacancy rates as an indication of housing availability.

To obtain a measure of fitness for use for this use case example, counts of counties, places, and tracts where the occupancy rate exceeds 90 percent in the MDF, but is below 90 percent in the CEF will be provided. Counts of where the error of occupancy rate is 2 percentage points to 5 percentage points and more than 5 percentage points will also be provided. [Use Case Table 11]

Additional Funding for Public Services

Numeric Thresholds

In this use case, a threshold has been established where once an area crosses that threshold, additional funds to meet the needs of the area are made available. A specific example is the provision of additional funds to hire additional police officers once an area exceeds a population of 50,000.

To obtain a measure of fitness for use for this use case example, counts of counties, place level geographies, and tracts where:

- total population equals or exceeds 50,000 in the MDF but is below 50,000 in the CEF
- total population is below 50,000 people in the MDF but equals or exceeds 50,000 people in the CEF
  [Use Case Table 12]

<u>Full Demographic and Housing Characteristics File (DHC) Variables Use Cases</u>

The Tenure and Relationship variables, planned for inclusion in the DHC, were not available in the 2010 Demonstration Data Products and will not be available until the DAS is fully scaled. Metrics for these variables are provided below.

Tenure

MAE, RMSE, MAPE, CV, and MALPE will be provided for the following tenure categories: Owned with a mortgage, Owned free and clear, and Rented for the following geographies [DHC Use Case Table 1.a-d]:

- All states
- Counties by size categories
- Incorporated places by size categories

Size categories for counties are:

- Less than 1,000 people
- 1,000 to 4,999 people
- 5,000 to 9,999 people
- 10,000 to 49,999 people
- 50,000 to 99,999 people
- 100,000 people or more

Size categories for incorporated places are:

- Less than 500 people
- 500 to 999 people
- 1,000 to 4,999 people
- 5,000 to 9,999 people
- 10,000 to 49,999 people
- 50,000 to 99,999 people
- 100,000 people or more

MAE and RMSE will be provided for the following tenure categories at the tract level: Owned with a mortgage, Owned free and clear, and Rented

For all geographies, secondary measures of outliers will be provided for tenure categories: Owned with a mortgage, Owned free and clear, and Rented. This measure will include counts of geographies where the absolute percent difference "exceeds 5 percent."

MAE, RMSE, MAPE, CV, and MALPE will also be provided for owner occupied and renter occupied by Householder Age for five age groups (15-24, 25-34, 35-54, 55-64, and 65 and over) [DHC Use Case Table 2a-c], owner occupied and renter occupied by Householder Hispanic Origin [DHC Use Case Table 3a-c], and owner occupied and renter occupied by Householder Race [DHC Use Case Table 4a-c] for the following geographies:

- All states
- Counties
- Incorporated places

For all geographies, a secondary measure of outliers will be provided to include counts of geographies where the absolute percent difference "exceeds 5 percent."

Detailed Vacancy Status
MAE, RMSE, MAPE, CV, and MALPE will be provided for counts of housing units by Detailed Vacancy Status [DHC Use Case Table 5a-c] for the following geographies:

- All states
- Counties by size categories
- Incorporated places by size categories

Size categories for counties are:

- Less than 1,000 people
- 1,000 to 4,999 people
- 5,000 to 9,999 people
- 10,000 to 49,999 people
- 50,000 to 99,999 people
- 100,000 people or more

Size categories for incorporated places are:

- Less than 500 people
- 500 to 999 people
- 1,000 to 4,999 people
- 5,000 to 9,999 people
- 10,000 to 49,999 people
- 50,000 to 99,999 people
- 100,000 people or more

MAE and RMSE will be provided for counts of housing units by Detailed Vacancy Status [DHC Use Case Table 5d] for the following geographies:

- Tracts

For all geographies, a secondary measure of outliers will be provided to include counts of geographies where the absolute percent difference "exceeds 5 percent."

An additional table will provide an outlier measure for tracts that shows the count of tracts where the largest vacancy type in the CEF has changed to another vacancy type in the MDF [DHC Use Case Table 6].

Household Size Categories
MAE, RMSE, MAPE, CV, and MALPE will be provided by the number of households by household size groupings (0 person household, 1 person household, 2 person household, 3 person household, 4 person household, 5 person household, 6 person household, and 7 or more person household) for the following geographies [DHC Use Case Table 7a-c]:

- All states
- Counties by size categories
- Incorporated places by size categories

Size categories for counties are:

- Less than 1,000 people

- 1,000 to 4,999 people
- 5,000 to 9,999 people
- 10,000 to 49,999 people
- 50,000 to 99,999 people
- 100,000 people or more

Size categories for incorporated places are:

- Less than 500 people
- 500 to 999 people
- 1,000 to 4,999 people
- 5,000 to 9,999 people
- 10,000 to 49,999 people
- 50,000 to 99,999 people
- 100,000 people or more

For all geographies, a secondary measure of outliers will be provided to include counts of geographies where the absolute percent difference “exceeds 5 percent.”

Outreach for Specific Household Types – Relationship Use Cases

Variable Sum Total

These use cases depend on the accuracy of the data for specific household types. For these types of use cases, fitness of use depends on being able to correctly identify a concentration of specific household types.

Data on multigenerational households are used for housing planning; funding distribution; support for elderly care; foster assistance for grandchildren being cared for by their grandparents; and for planning economic development, housing services, transportation, community development, and long-range planning.

MAE, RMSE, MAPE, CV, and MALPE will be provided for counts of households by presence of people 65 years and over living alone [DHC Use Case Table 8a-c] and presence of multigenerational households by Hispanic Origin of householder and by race of householder [DHC Use Case Table 9a-c; 10a-c] for the following geographies:

- All states
- Counties by size categories
- Incorporated places by size categories

Size categories for counties are:

- Less than 1,000 people
- 1,000 to 4,999 people
- 5,000 to 9,999 people
- 10,000 to 49,999 people
- 50,000 to 99,999 people
- 100,000 people or more

Size categories for incorporated places are:

- Less than 500 people
- 500 to 999 people
- 1,000 to 4,999 people
- 5,000 to 9,999 people
- 10,000 to 49,999 people
- 50,000 to 99,999 people
- 100,000 people or more

For all geographies, a secondary measure of outliers will be provided to include counts of geographies where the absolute percent difference "exceeds 5 percent."

Disaggregated data on single-parent households are critical to measuring the conditions, well-being, and progress of the many diverse communities in the United States. Without access to these data, public officials cannot effectively address the needs of many smaller, and often most vulnerable, population groups. Data are used for program management; projections concerning community needs and participation in public programs; and planning for economic development, housing services, transportation, community development, long-range plans, and child care services.

MAE, RMSE, MAPE, CV, and MALPE will be provided for household type by presence of own children under 18 years old [DHC Use Case Table 11a-c] and count of households by presence of own children under 6 years old [DHC Use Case Table 12a-c] for the following geographies:

- All states
- Counties by size categories
- Incorporated places by size categories

Size categories for counties are:

- Less than 1,000 people
- 1,000 to 4,999 people
- 5,000 to 9,999 people
- 10,000 to 49,999 people
- 50,000 to 99,999 people
- 100,000 people or more

Size categories for incorporated places are:

- Less than 500 people
- 500 to 999 people
- 1,000 to 4,999 people
- 5,000 to 9,999 people
- 10,000 to 49,999 people
- 50,000 to 99,999 people
- 100,000 people or more

For all geographies, a secondary measure of outliers will be provided to include counts of geographies where the absolute percent difference “exceeds 5 percent.”

Data on the number of married and unmarried-partner households are used in strategic planning; local and regional planning of housing; land use; transportation; economic development; Community Development Block Grant (CDBG) programs; grant applications; working towards equitable communities and opportunities; forecasting trends of population growth; and analyzing shifting demographics. The population in a same-sex relationship (married or unmarried partner) is a relatively small group. If the quality of these data are undermined, public officials cannot effectively address the needs of many smaller, and often most vulnerable, population groups.

MAE, RMSE, MAPE, CV, and MALPE will be provided for opposite-sex and same-sex married couples and unmarried partner households by Hispanic origin of householder [DHC Use Case Table 13a-c] and by race of householder [DHC Use Case Table 14a-c] for the following geographies:

- All states
- Counties
- Incorporated places

For all geographies, a secondary measure of outliers will be provided to include counts of geographies where the absolute percent difference “exceeds 5 percent.”

Policy makers use information about the number of children by whether they are living with a biological, adoptive, or stepparent, or with a grandparent to estimate how many people or households are eligible for particular government programs. Adopted children and foster children are a particular policy concern, as well as grandchildren who can also be foster or adopted children. There are various government programs that apply to these children and their families, depending on the nature of the foster arrangement or method of adoption, whether international, private, or through the state-run foster care system.

MAE, RMSE, MAPE, CV, and MALPE will be provided for the count of people in each relationship category, including child and grandchild categories, [DHC Use Case Table 15a-c] for the following geographies:

- All states
- Counties
- Incorporated places

For all geographies, a secondary measure of outliers will be provided to include counts of geographies where the absolute percent difference “exceeds 5 percent.”

Impossible or Improbable Results Use Cases

Impossible or improbable results were a major finding from analyses of the 2010 Demonstration Product and can raise questions about the validity of the Census counts. The following tables are dedicated to identifying impossible and improbable results. In many cases, the measures provided in this section reflect those provided to the Census Bureau by external stakeholders. The improbable results should be considered relative to the same measures from the CEF.

State Total Population Invariant

Total population at the state level is invariant so a measure of accuracy is not needed. Instead, a table showing the count of states where the total population in the CEF is different from the MDF will be provided at the state level as a confirmation that state total population did not change. [Inconsistency Table 1]

Impossible Results

A count of blocks where the following impossible conditions are met will also be provided. These conditions are meant to represent impossible values that were seen in the 2010 Demonstration Product and are not representative of all potential impossible values.

- Review of the demonstration product revealed population, household size, and household counts that when considered together represented impossible values. This was due to inconsistencies between the person file, which contains person information, and the housing unit file, which contains housing information; these inconsistencies resulted from applying disclosure protections to each of these file separately. The following two measures are meant to show the extent of these inconsistencies. A count of tracts where households from the person file outnumber people when the count of people is derived from the household size variable will be provided. Even though the household size variable includes a "Size +7" category, by assuming those households all have the smallest size of 7, an approximation of the population count can be obtained. This value can be compared to the population total from the person file. A count of the number of tracts where the population total is less than the population derived from the household size variable will also be provided. [Inconsistency Table 2]
- Household population is less than the number of occupied housing units. (Universe: Blocks with at least 1 occupied housing unit) [Inconsistency Table 3]
- Household population is equal to or greater than 1 while the number of occupied housing units is 0 (Universe: Blocks with at least 1 person in households) [Inconsistency Table 4]
- The number of occupied housing units is equal to or greater than 1 while the household population is 0 (Universe: Blocks with at least 1 occupied housing unit) [Inconsistency Table 5]
- Tracts with more householders of a certain race than population of that race (Universe: Tracts with at least 1 person and no GQ population) [Inconsistency Table 6]
- Tracts with more householders who are Hispanic or Latino than population that are Hispanic or Latino (Universe: Tracts with at least 1 person and no GQ population) [Inconsistency Table 7]
- Tracts with more householders of a certain age than population of that age (Universe: Tracts with at least 1 person and no GQ population) [Inconsistency Table 8]
- Tracts with more householders who are female than population that are female (Universe: Tracts with at least 1 person and no GQ population) [Inconsistency Table 9]
- Tracts with more households with children under 18 years old than people under 18 years old (Universe: Tracts with at least 1 person and no GQ population) [Inconsistency Table 10]
- Tracts with more opposite-sex married couple households than males (Universe: Tracts with at least 1 person and no GQ population) [Inconsistency Table 11]
- Tracts with more opposite-sex married couple households than females (Universe: Tracts with at least 1 person and no GQ population) [Inconsistency Table 12]
- Tracts with more married couple households than population age 15 years and over multiplied by two (Universe: Tracts with at least 1 person and no GQ population) [Inconsistency Table 13]

Improbable Results

A count of geographies where the following improbable conditions are met will also be provided. These conditions are meant to represent possible but improbable values that were seen in the 2010 Demonstration Product and is not representative of all potential impossible values.

- Counties and tracts with at least 5 children under age 5 and no women age 18 through 44 (Universe: Counties and tracts with at least 5 children under age 5) [Inconsistency Table 14]
- Counties and tracts with at least 5 children under age 5 of a certain major race group and no women age 18 through 44 of the same race group (Universe: Counties and tracts with at least 5 children under age 5 of a certain major race group) [Inconsistency Table 15]
- Tracts with at least 5 people and all of the same sex (Universe: Tracts with at least 5 people) [Inconsistency Table 16]
- Tracts with at least one of the single years of age between 0 and 17 by sex has a zero count (Universe: Tracts with 200 or more 0-17 year olds) [Inconsistency Table 17]
- Blocks with population all 17 or younger (Universe: Blocks with at least 1 person and no GQ population) [Inconsistency Table 18]
- Blocks with persons per household greater than 10 (Universe: Blocks with at least 1 occupied housing unit) [Inconsistency Table 19]
- Counties and tracts where median age of the men is significantly different (equal to or greater than 20 years) from the median age of women, by major race group (Universe: Counties and tracts with at least 5 males and 5 females of major race group) [Inconsistency Table 20]
- Tracts with 100% of the population in “adult” group quarters with population under 18 years (Universe: Tracts with 100% of the population in “adult” group quarters) [Inconsistency Table 21]
- Blocks where the occupancy is 100 percent in the MDF but not the CEF [Inconsistency Table 22]
- Blocks where the occupancy is 0 percent in the MDF but not in the CEF [Inconsistency Table 23]

Appendix: Measures of Accuracy

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- Mean Absolute Percent Error (MAPE)
- Coefficient of Variation (CV)
- Mean Algebraic Percent Error (MALPE)
- Root Mean Squared Error
- Percent Difference Thresholds
- Total Absolute Error of Shares

Mean Absolute Error (MAE) = $(\Sigma\ (|\text{MDF} - \text{CEF}|))/N$
MAE takes the absolute value of the difference between the MDF and the CEF value for each evaluation geography, sums them, and divides by the number of evaluation geographies. The goal is to provide an easy to interpret measure of the numeric error.

Root Mean Squared Error = $\text{SQRT}(\Sigma\ ((\text{MDF} - \text{CEF})^2)/N)$
This measure squares the difference between the MDF and the CEF number for each evaluation geography, sums these values across evaluation geographies, divides by the number of evaluation geographies, and finds the square root of this value. It presents an alternative measure that places greater emphasis on large numeric errors versus mean absolute errors.

Mean Absolute Percent Error (MAPE) = $((\Sigma\ ((|\text{MDF} - \text{CEF}|)/\ \text{CEF}))/N)*100$
MAPE takes the absolute value of the difference between the MDF and the CEF value for each evaluation geography, divides that by each respective CEF value, sums them, divides by the number of evaluation geographies, and multiplies the result by 100. The goal is to provide an easy to interpret relative measure of error. This is one of the most commonly used measures for assessing the accuracy of a series of population estimates.

Coefficient of Variation = $(\text{RMSE}/(\Sigma\ (\text{CEF})/N))*100$
This measure restates the RMSE as a percentage of the average statistic in the geography.

Mean Algebraic Percent Error (MALPE) = $((\Sigma((\text{MDF} - \text{CEF})/\text{CEF}))/N)*100$
MALPE takes the difference between the MDF and the CEF value for each evaluation geography, divides that by each respective census value, sums them, divides by the number of evaluation geographies, and multiplies the result by 100. Its purpose is to identify systematic bias and provide an alternative for a relative measure of error.

Percent Difference Thresholds = Number of absolute percent differences above a certain threshold
Unlike the other measures, Percent Difference Thresholds is a numeric value that relies upon an arbitrarily set threshold (e.g., 5 and 10 percent). In short, the absolute percent difference is computed by dividing the absolute difference between the MDF and CEF value for a given area by the CEF value for that area and multiplying by 100. The end measure simply represents a count of how many evaluation geographies in the summary area exceeded a particular threshold in their absolute percent difference of the estimate. It provides an intuitive measure of the distribution of differences.

Total Absolute Error of Shares = Σ|((MDF/ΣMDF) – (CEF/ΣCEF))|

This measure finds the proportion of each MDF value to the total MDF value for the summary geography and subtracts the proportion of the CEF value to the total CEF value for the summary geography. The absolute value of these proportional differences across evaluation geographies is then summed to the summary geography level. The goal is to provide a measure of the distributional error in the MDF shares.